NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

FABRICATING SYNTHETIC DATA IN SUPPORT OF TRAINING FOR DOMESTIC TERRORIST ACTIVITY DATA MINING RESEARCH

by

Stephen Lavelle

September 2010

Thesis Advisor: Simson Garfinkel
Second Reader: George Dinolt

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** 24 September 2010	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE Fabricating Synthetic Data in Support of Training for Domestic Terrorist Activity Data Mining Research		**5. FUNDING NUMBERS**	
6. AUTHOR(S) Stephen Lavelle			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.		**12b. DISTRIBUTION CODE**	
13. ABSTRACT Data mining is a mature technology, widespread in both government and industry. The proliferation of data storage in public and private sectors has provided more information than can be expediently processed. Data mining provides a means to extract meaningful conclusions from this growing store of data. In the interests of countering criminal and terrorist activity, data mining has become a focus of law enforcement and government agencies. The use of databases containing information on persons may conflict with privacy rights and laws. Gathering public awareness of government data mining programs and databases has been accompanied with concern and investigation of these programs. Following a review of data mining and privacy issues, in 2008 the National Research Council (NRC) recommended any training in development of data mining programs involving personal data be conducted using synthesized data. This thesis seeks to present an underlying discussion of these issues, to include data mining use, a simple data synthesis model for analysis to support the validity of the NRC recommendation, and the associated difficulties encountered in the process. Included is an analysis of the inherent difficulty in creating realistic and useful data.			
14. SUBJECT TERMS Data Synthesis, Data Mining, Counterterrorism			**15. NUMBER OF PAGES** 105
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited.

FABRICATING SYNTHETIC DATA IN SUPPORT OF TRAINING FOR DOMESTIC TERRORIST ACTIVITY DATA MINING RESEARCH

Stephen J. Lavelle
Major, United States Marine Corps
B.S., State University of New York, 1994

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
September 2010

Author: Stephen J. Lavelle

Approved by: Simson Garfinkel
Thesis Advisor

George Dinolt
Second Reader

Peter Denning
Chairman, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Data mining is a mature technology, widespread in both government and industry. The proliferation of data storage in public and private sectors has provided more information than can be expediently processed. Data mining provides a means to extract meaningful conclusions from this growing store of data. In the interests of countering criminal and terrorist activity, data mining has become a focus of law enforcement and government agencies. The use of databases containing information on persons may conflict with privacy rights and laws. Gathering public awareness of government data mining programs and databases has been accompanied with concern and investigation of these programs. Following a review of data mining and privacy issues, in 2008 the National Research Council (NRC) recommended any training in development of data mining programs involving personal data be conducted using synthesized data. This thesis seeks to present an underlying discussion of these issues, to include data mining use, a simple data synthesis model for analysis to support the validity of the NRC recommendation, and the associated difficulties encountered in the process. Included is an analysis of the inherent difficulty in creating realistic and useful data.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. INTRODUCTION 1
A. MOTIVATION 2
B. SCOPE OF THESIS 3

II. PRIOR AND RELATED WORK 5
A. USE OF SYNTHETIC DATA 5
B. DATA SANITIZATION/ANONYMIZATION 8
C. USE OF REAL DATA 9
D. CONCLUSION 10

III. DATA MINING FOR COUNTERTERRORISM IN THE U.S. 13
A. BACKGROUND 13
B. DATA MINING: LINK ANALYSIS VS. PATTERN MATCHING 14
1. Link Analysis 15
2. Pattern-Based Data Mining 16
C. THE EMERGENCE OF DATA MINING AS A COUNTERTERRORISM TOOL 17
1. 9/11 Attacks and Resulting Actions 18
2. Public Awareness of Government Activity 19
3. Data Mining Reporting Act of 2007 22
4. Lack of Closure on Programs 24
D. DHS REPORTS ON DATA MINING PROGRAMS (2008-2009) 26
1. Automated Targeting System (ATS) 27
2. Data Analysis and Research for Trade Transparency System (DARTTS) 28
3. The Freight Assessment System (FAS) 28
E. FOCUS ON A SPECIFIC PROGRAM: ATS IN DETAIL 29
1. ATS Passenger Module (ATS-P) 30
2. ATS Inbound Module (ATS-I) 31
F. THE NATIONAL RESEARCH COUNCIL RECOMMENDATION 31
G. CONCLUSION 32

IV. DATA MINING RESEARCH AND SYNTHETIC DATA 35
A. PURPOSE OF SYNTHETIC DATA 35
B. DATA CHARCTERISTICS 37
1. Realism 37
2. Usefulness 39
C. DATA SYNTHESIS; APPROACHES TO MODELING 39
1. Manual Creation 40
2. Automated Direct Creation of Realistic Data 40
3. Generative Simulation 41
4. Comparison of Approaches 42
D. DATA SYNTHESIS; ADDITIONAL CONSIDERATIONS 44
1. Creating Names 44

2. Protecting Data 46
V. DATA MODALITIES 47
A. TYPES OF DATA 47
1. Record-Based 47
2. Photographic Data 49
3. Biometric Data 50
4. Narrative Data 51
5. Free Format Data 52
B. SIMPLIFIED TEST DATA 53
VI. A DATA SYNTHESIS EXERCISE 55
A. REPRODUCING THE DATA 55
1. Deconstructing the ATS 55
2. A Realistic Data Group to Populate 57
3. Available Statistics to Support Realism 58
4. Secondary Database 58
B. MODEL SPECIFICATION 59
1. Inputs 60
2. Outputs 63
3. Treatment of Inputs/Record Creation 63
4. Deliberate Introduction of Error 66
5. Control of Duplication 67
6. Limitations of This Model 68
VII. FURTHER CONSIDERATIONS AND FUTURE WORK 71
A. CHALLENGES TO CREATING SYNTHETIC DATA 71
1. Effort of Human Designer 71
2. Perceived Privacy of Data 72
3. Human/Designer Bias 73
B. FUTURE WORK 74
C. CONCLUSIONS 76
APPENDIX A: DHS I-901 FORM 77
APPENDIX B: SAMPLE STUDENT STATISTICS (CSU LONG BEACH) 79
APPENDIX C: SAMPLE CALIFORNIA DRIVER'S LICENSE APPLICATION (FORM DL44) 81
LIST OF REFERENCES 83
INITIAL DISTRIBUTION LIST 87

LIST OF FIGURES

Figure 1. Model Overview, Automated Direct Creation.....61
Figure 2. Sample “seed” surname file.csv.....62
Figure 3. Class:Person.....64

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1. Operational or Planned Data Mining Programs 21

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ATS	Automated Targeting System
CBP	Customs and Border Protection
DARTTS	Data Analysis and Research for Trade Transparency System
DHS	Department of Homeland Security
DOE	Department of Education
DOJ	Department of justice
FAS	Freight Assessment System
GAO	Governmental Accounting Office
ICE	Immigration and Customs Enforcement
IRS	Internal Revenue Service
MATRIX	Multi-State Anti-Terrorism Information Exchange
NRC	National Research Council
NSEERS	National Security Entry Exit Registration System
PII	Personal Identifying Information
SEVIS	Student and Exchange Visitor Information System
TECS	Treasury Enforcement Communication System
TIA	Total Information Awareness
TSA	Transportation Security Administration
US-VISIT	United States Visitor and Immigrant Status Indicator Technology

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I extend my profound gratitude and appreciation to Dr. Simson Garfinkel, whose unwavering dedication to assist me in formulating this thesis was invaluable.

I appreciate the insightful review of Dr. George Dinolt, whose experience and wisdom was instrumental to the refinement and completion of the thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This thesis seeks to address the problem of creating realistic and useful data in support of data mining programs designed to the threat of domestic terrorism.

Since the attacks of 2001, the federal government has implemented many data mining programs in support of counterterrorism. Information on the efficacy of these programs and their methodology is very limited. Many of the data mining programs are classified, creating additional problems for meaningful public discussion. Because many such programs necessarily involve the use of personally identifiable information (PII), concern has arisen over the implementation of these programs and what impact they have on the privacy of U.S. citizens. Improving, expanding, or modifying programs presents a challenge without a useful and realistic body of data from which to test the systems.

Real data representing PII of U.S. citizens would present one potential test data set for the research and testing of data mining approaches. Such data is attractive because ultimately it is largely this data that will be used as input to these programs. These bodies of data (e.g., individual Social Security numbers, tax data, etc.), however, are governed by privacy laws and regulations as to the purposes and disclosure of their use to affected individuals.

Because the use of real data to develop or test data mining programs using PII is problematic, a possible

alternative is the use of synthetic data. Such data must be created, however, which remains a challenge. Creating data that mirrors real data is a hard problem. Nonetheless, given the requirement and motivation by law enforcement for data mining programs utilizing PII, a real need for simulated data exists.

A. MOTIVATION

A fundamental part of counterterrorism activities is detecting and preventing terrorism. Those who would commit terrorism frequently first decide to commit such acts and subsequently make plans. Data mining technology may present a useful tool to detect such planning. Analysis of circumstances surrounding the attacks of September 2001 revealed numerous correlations in the patterns of the attackers before the attacks occurred. The use of data mining technology in such a circumstance could have allowed investigators to detect and investigate the perpetrators before the attacks, which presumably might have prevented the attacks from occurring.

Planning to commit a terrorist act is in itself a punishable crime, if the intended act would involve destruction of a public transportation system[1]. Thus, if the September 2001 attackers had been caught in the act of planning the attacks, they could have been tried, convicted and imprisoned even if the attacks had not been carried out.

[1] 18 U.S.C. Sec. 2332.

In 2008 the National Research Council of the National Academies released its report "Protecting Individual Privacy in the Struggle Against Terrorists" (National Research Council, 2008), which analyzed federal data mining efforts in light of growing privacy concerns. The NRC report concluded "…high priority should be given to developing programs to detect intended attacks before they occur so that there is a chance of preventing them."

A problem with this approach, however, is that in the planning phase potential terrorists have not yet committed a violent act, and are otherwise co-mingled with regular citizens in the routines of daily life. How can a program detect and cull the terrorists from the innocent citizens without studying the behavior or relationships of both, and thereby invade the privacy of the innocent?

Given these problems, the NRC concluded as one of its key recommendations that any data mining programs intending to use the PII of U.S. citizens should first be created and tested using synthetic data in order to protect the privacy of citizens.

B. SCOPE OF THESIS

This thesis will explore the combined issues of data mining and privacy within the framework presented in the recommendations of the NRC. Specifically, the problem of creating realistic and useful synthesized (i.e., fake) data will be addressed. A study of data will be conducted exploring the complexity of manufacturing realistic and useful data. Existing federal data mining programs will be

presented in order to grasp the measure and make-up of the data used in such programs. Finally, an analysis of the inherent obstacles to creating synthetic data will be presented in light of the NRC recommendation.

For the purposes of this analysis, the concept of data "aggregation" or "fusion" may be implied in the discussion of some data mining programs or functions as either a subset or result of data mining actions, though technically they may be viewed as separate functional areas.

II. PRIOR AND RELATED WORK

Data mining efforts and research have attempted to mitigate privacy concerns with work in three areas; synthesized data, data sanitization/anonymization, and real data not subject to privacy concerns. The following three sections will explore these types of data in more detail.

A. USE OF SYNTHETIC DATA

Existing research regarding the creation of synthesized data is found in several areas, to include fraud detection, medical use, and PII. The latter of these three is the focus of this study. Other efforts of interest include creation of data for the purposes of population and urban planning, evaluation of computer security, and data generation for digital forensics training.

Data generation modules in support of counterterrorism research for data mining have been developed by the University of California, Riverside in conjunction with Lucent technologies. An initial design created a data generation module to create person names and credit card information, to include transaction information (Jeske, et al., 2005). The application creates names from lists of first names and surnames, social security numbers from the Social Security Administration's deceased list, and credit card numbers based on the rules associated with actual credit card company number pools. Other attributes, such as age, income, gender, etc. were created and associated

using three types of algorithms to generate credit card transaction information. A second iteration later added a new shipping container (related to importing goods) information application (Jeske, Lin, Rendon, Xiao, & Samadi, 2006).

A team from the University of Arkansas addressed an issue of the need for large "industrial" data sets in order to properly test data mining systems, and presented a model using parallelism (i.e., data generation across multiple processors) (Hoag & Thompson, 2007).

The open source Urban Sim project provides a model for urban development for land use, environment and transportation planning (Waddell, 2002). According to the article "Urban Sim: Modeling Urban Development for Land Use, Transportation and Environmental Planning," the project encompasses a wide capability range including city and population modeling based on census, land parcel and business data, among others. Included in the article is a more detailed comparison of Urban Sim with three related and/or previous efforts, many of which faced the same issue of synthetic data production. Urban Sim does not make use of PII (real or synthesized), though presumably it could be expanded to do so.

Within the field of computer security, another example of synthesized data includes that created by Lincoln Labs to test the Defense Advanced Research Projects Agency (DARPA) developed Intrusion Detection systems in 1998-99 (Lee, Stolfo, & Mok, 1999). Given the sensitivity of real data, the project created sets of attack data and routine

organization data "noise." McHugh offers insights and criticism regarding the lack of realism contained in the data sets created in the DARPA project (McHugh, 2000).

A further example of synthetic data is in support of digital forensics investigator training. The Forensic Image Generator project of the University of Mannheim seeks to produce realistic digital images for student analysis (Moch & Freiling, 2009). Using instructor generated scripts and the Python programming language, the Generator produces file system images that attempt to mirror real files that might be generated on a computer system by a user. As of 2009 the Generator was still in its prototype stage. Like data mining for counterterrorism, the motivation for this project also stems from the need to avoid using real data to protect privacy.

A final example is presented in the computer gaming world. The "Sim" products from Electronic Arts allow for the fabrication of worlds and people. The version *Sim City* offers gamers the ability to create and manage synthetic cities over time. *The Sims* versions allow for the creation and manipulation of individual people. In terms of application to data synthesis for data mining for counterterrorism, neither of these products is useful. *Sim City* has a level of resolution of detail too low, and *The Sims* has a level too high. Perhaps more importantly, neither system allows for a principled simulation from grounded assumptions, or the automated collection of simulation data.

B. DATA SANITIZATION/ANONYMIZATION

Efforts and proposals to utilize real data that has been sanitized to remove Personally Identifiable Information (PII) have not been fully successful under public and academic scrutiny. Research programs, such as the Total Information Awareness (TIA) program, proposed a combination of methods. The methods proposed included privacy appliances (cryptographic mechanisms inserted between human users and databases), methods of data hiding (revealing limited record information not completely displaying all elements of PII), selective revelation (revealing greater amounts of PII as data qualified for more potential malicious activity), and effective oversight through independent audit (Popp & Poindexter, 2006). Opponents and proponents of data sanitization approaches argue the validity of privacy preservation. While Popp and Poindexter (2006) may present privacy preservation approaches as effective or acceptable means to hiding PII, others argue that the mere existence of databases containing PII (and the fusion of such databases into new data records) potentially creates a violation of individual privacy, especially if it leads to unwanted surveillance (Garfinkel, 2006). The existence of these databases may be unknown to the individuals whose information they contain. Several questions arise from this discussion. A fundamental question is, if such databases contain PII and exist in some electronic media but are never actually seen by analysts, has a breach of privacy been committed? If such records are created, does the individual have a

reasonable expectation that it will be securely stored and/or properly deleted at some point in time?

A second case is the Netflix Prize wherein the NetFlix movie rental-by-mail company released sanitized records of more than 500 million individuals' movie preferences (Bennett & Lanning, 2007). Using data perturbation techniques and other data available on Internet movie sites, the dataset was "re-identified," offering compromised access to actual individual identities and political affiliations (Narayanan & Shmatikov, 2008).

Similar to the NetFlix re-identification problem, in 2006 the New York Times reported that the internet provider America Online (AOL) released a large dataset containing the search data of some 657,000 users (M. Barbaro & T. Zeller, 2006). The 20 million records displayed specific search terms entered by users with each individual user "anonymized" by an AOL provided user number. Users sued after release of the data set resulted in incidences of re-identification and general privacy concerns (CNN Money, 2006).

C. USE OF REAL DATA

In other situations, it has been shown that real data not subject to privacy concerns can be used to trace malicious activities. This category is useful not only in demonstrating the usefulness of data mining in general, but also in segregating that portion of data available for analysis without regard to privacy concerns. Such data includes criminal records. The COPLINK research project of

the University of Arizona has demonstrated success discovering deceptive aliases of criminals using real data from the Tucson Police Department (Wang, Chen, & Atabakhsh, 2004). Using records of known criminals and records using known aliases from actual Police records, the project was able to correctly identify criminals at a success rate greater than 95%. Other areas of study under the same project showing promising results include criminal network analysis, police report entity extraction, and authorship analysis in cybercrime.

Other non-private data used in data mining includes international trade data. Using data published by the U.S. Department of Commerce, Bureau of Census, and contained in the U.S. Merchandise trade database, potential money laundering and terrorist financing through overvaluing or undervaluing imports and exports was illustrated in "Detecting Money Laundering and Terrorist Financing via Data Mining" (Zdanowicz, 2004).

D. CONCLUSION

While data mining continues to evolve as a mature technology, evidence of a protracted effort to develop and use synthetic data for data mining research has not been found. Some research into modeling medical or population data exists, but there are few programs that model and manufacture realistic personally identifiable information (PII), or which do so on a broad scope. Moreover, in its recommendation to use synthetic data for the development of data mining programs, the National Research Council did not

offer any guidance on how such data might be modeled or created (National Research Council, 2008).

THIS PAGE INTENTIONALLY LEFT BLANK

III. DATA MINING FOR COUNTERTERRORISM IN THE U.S.

A. BACKGROUND

Data mining technology has seen considerable advancement in capability and use over the past two decades. (Fayyad & Uthurusamy, 2002) offers a simple definition of data mining as "the identification of interesting structure in data." (Apte, 1997) offers yet another definition: "…the process by which accurate and previously unknown information is extracted from large volumes of data." As data storage capability has rapidly increased across numerous functional areas, vast data warehouses have become available for the practice of data mining (Fayyad & Uthurusamy, 2002). Data mining is commonly used for commercial and medical purposes. Commercial data mining might seek to aid business leaders in pursuing targeted customer bases while medical data mining seeks to maximize patient health care or discover trends in ailments. Data mining can also be used to search for evidence of criminal activity and its perpetrators. This kind of data mining is generally conducted by law enforcement agencies. In this capacity data mining offers considerable opportunity to prevent crime or identify potential criminals or terrorists.

The gravity of the attacks against the United States in September 2001 focused law enforcement agencies on the vital task of preventing future attacks against innocent populations. Given the increasing reliance on technology and vast data available via the global information grid,

data mining has emerged as a widely-used method to detect criminal patterns and uncover potentially malicious social networks (Governmental Accounting Office, 2004). As worldwide data storage capacity has increasingly exceeded computing capability, the need for efficient extraction of relevant but buried data and follow-on analysis becomes evident (Fayyad & Uthurusamy, 2002).

As a counterterrorism tool, data mining faces challenges. Data potentially containing information on terrorists is necessarily mixed with the data of non-terrorists. Public awareness of federal activities, including the Information Awareness Office's Total Information Awareness (TIA) program in 2002 and alleged National Security Agency use of public phone records in 2003, resulted in Congressional hearings and widespread criticism of data mining programs in relation to privacy concerns (Garfinkel, 2006). Additionally, the need for useful database records (i.e., data in useful format), presumably gathered from numerous active sources both public and private, would require significant management and maintenance.

B. DATA MINING: LINK ANALYSIS VS. PATTERN MATCHING

Data mining for terrorism could implement numerous technologies and methodologies to achieve its goal. A complete survey of data mining technology is outside the scope of this paper; however, discussion of two main categories of data mining, link analysis and pattern matching, is relevant to the focus of this paper.

1. Link Analysis

Link analysis, sometimes referred to as 'subject-based' data mining, offers a simple but effective form of data mining and underscores the usefulness of data mining given large data sets. With link analysis, a known data item of interest can be linked to other items based on actual events in reality. For instance, if a certain street address is known to be a criminal "hide-out," an individual having the same residence could be linked through an address-matching algorithm to the criminal activity. Such links could easily be made on a small scale by an investigator following traditional investigative methods. Given the large data sets available containing such records, link analysis data mining presents a useful automated approach to the procedure.

Link analysis suffers from several technical problems including errors in data and differences in schemas. For example, the data entry representing an individual who resides at "123 Baker Street" might actually be recorded as residing at "124 Baker Street" or "123 Bokor Street." A simple link analysis data mining program might fail to match two separate records based on these errors. Entries might also differ in other ways. One database might display an address as "123 Baker Street" while another might present the same address as "0123 Baker Street," "#123 Baker Street," or "123 BAKER STREET".

Two or more databases that use different schemas to organize their data and used as input to a data mining program can create problems for link analysis. For

example, one data base might present an individual's address in three fields; number, street name, and street descriptor (e.g., "Avenue"). Another database might simply have one field "Address" containing all of the data." Numerous other variations in data presentation could result, requiring careful design of programs to allow for differences. For example, one database might represent the address with street name and number, while another might list the real property data having book number or plot number.

DeRosa (2004) provides an illustration of the potential of link analysis, demonstrating the links (termed "degrees of separation") between all of the September 2001 attackers beginning with the three known suspects. Each of the attackers could be linked to the original three known terrorist suspects.

2. Pattern-Based Data Mining

A second major category of data mining technology is pattern-based data mining. This method seeks to detect malicious activity (in the case of data mining for counterterrorism) by discovering patterns in behavior. In this case, unlike link analysis, an individual or specific place may not be known to investigators *a priori*. Instead, a specific pattern of behavior (such as the example of purchasing large amounts of fertilizer and renting a truck) may be used as a filter. The identity of individuals matching this pattern of activity may then be culled from large amounts of data and subsequently targeted for further investigation.

Jonas and Harper (2006) point out that pattern-based (or "predictive") data mining suffers from several inherent problems. First, such programs invariably result in numerous "false positives." In other words, if a program resulted in 100 potential leads to criminal activity, only 5% might actually be valid. Researching each lead of potential criminal activity is very costly in terms of resources; because of the high false positive rate, much investigative effort is wasted. Moreover, the false leads, besides being wasteful, results in innocent people being surveilled.

Jonas and Harper (2006) also note that predicting a pattern is easy if it is already known, but in reality terrorists may form new plots and methods that have not yet been used in order to achieve surprise. Unless such patterns can be 'guessed' by authorities prior to an attack, pattern-based data mining methods would not necessarily be helpful.

C. THE EMERGENCE OF DATA MINING AS A COUNTERTERRORISM TOOL

Analysis of past terrorist attacks has shown that data mining technology would have been useful in detecting some kind of terrorist planning and activities. While some dispute the efficacy of such data mining technologies as pattern based (or predictive) data mining, simple link analysis could have detected the events leading to many of these crimes (Jonas & Harper, 2006).

1. 9/11 Attacks and Resulting Actions

Leading up to September 2001, the attackers built a significant data trail of evidence related to the planning of the attack. Some of the attackers' names were known to authorities as possible terrorists linked to the USS Cole bombing and other known terrorists. Other 9/11 attackers left address records which matched those of the terrorists known to authorities. They obtained driver's licenses and held P.O. boxes in their own names, and maintained personal associations with the other known terrorists. The 9/11 commission report concluded that "by pursuing the leads available to it at the time, the government might have derailed the plan" (National Commission on Terrorist Attacks Upon the United States Washington DC, 2004).

The revelations of these correlations spurred the creation of numerous government data mining programs. A survey of many of these programs is included in "Total Information Awareness and Beyond: The Dangers of Using Data Mining Technology to Prevent Terrorism" (Anderson, 2003).

In the cases of both the 1993 World Trade Center bombing and the 1995 Oklahoma City bombing, similar actions perpetrated by the bombers could have potentially been detected by pattern-based data mining. In order to commit the bombings, the terrorists left a data trail from renting large trucks and buying large quantities of fertilizers containing nitrates. In both cases, the subjects were not agricultural workers. These events, when correlated, offer investigators insight into potential terrorist activities.

The ability to correlate such events presents an effective case for the use of data mining. As previously mentioned however, pattern-based data mining relies upon some previous knowledge or intuition of a pattern to search for.

2. Public Awareness of Government Activity

Following the September 2001 attacks, the public and the Congress increasingly became aware of governmental activities with regard to data mining, and their impact on privacy. In November 2002, the New York Times reported the existence of the Total Information Awareness (TIA) research program undertaken by the Defense Research Project Agency (DARPA). An opinion editorial by William Safire (2002) offered the following:

> If the Homeland Security Act is not amended before passage, here is what will happen to you:
>
> Every purchase you make with a credit card, every magazine subscription you buy and medical prescription you fill, every Web site you visit and e-mail you send or receive, every academic grade you receive, every bank deposit you make, every trip you book and every event you attend -- all these transactions and communications will go into what the Defense Department describes as "a virtual, centralized grand database."

Followed by public outcry and Congressional inquiry over privacy concerns, TIA was de-funded by Congress in 2003. Similarly, the FBI's Multi-State Anti-Terrorism Information Exchange (MATRIX), which focused on law-enforcement fusion among states and "…proposed to combine public records and private record data from multiple databases with data

analysis tools," failed by March 2004 due to several participating States withdrawing participation (Coney, 2007).

A third program, the Computer Assisted Prescreening System (CAPPS II), "…used computer-generated profiles to select passengers for additional screening…" (Dycus, Berney, Banks, & Raven-Hansen, 2007). CAPPS II was de-funded by Congress in 2005 due to similar concerns over the sharing of private information. While these programs were failing, other programs were being developed or planned by numerous government agencies.

In 2004, the Governmental Accounting Office (GAO) released its report on federal agencies' use of data mining for any purpose (Governmental Accounting Office, 2004). The GAO reported that of 128 federal agencies surveyed in the report, 52 were using or planning to use data mining in some form. Of these agencies, a total of 131 operational programs were reported, with the intent to begin 68 others. Of these 199 programs, 122 were reported to involve the use of personal data. It further revealed that 14 of 199 existing or planned programs were aimed at gathering information on potential terrorist activity. Of these programs, 10 utilized PII. The agencies ranged in functional category, from the Departments of Justice and Homeland Security to the Department of Education. The data used for these programs was resourced from both the public and private (commercial) sectors. Information available included data such as credit card transaction records and credit reports.

In its 2008 report on data mining, the Department of Homeland Security (DHS) reported three separate federal agencies conducting data mining programs to prevent terrorism, including the Transportation Security Administration (TSA), U.S. Customs and Border Protection (CBP), and Immigration and Customs Enforcement (ICE) (Department of Homeland Security, 2008). The three programs reported on were categorized from others as those that used personal data.

Given that classified data mining programs have been publicly acknowledged (Office of the Director of National Intelligence, 2009), and that the last GAO report on data mining from 2004 is now quite dated, it is uncertain how many programs currently utilize data mining technology.

The most comprehensive publicly available listing of federal programs, the 2004 GAO report, gives only very brief overviews in a survey fashion of 128 planned or operational programs across 52 federal departments. Ten of these programs were focused on counterterrorism and utilized PII as input (see FIGURE (1)). In 2005, the GAO did not repeat the survey, but focused instead on just five of the programs in the 2004 report, exploring their compliance with existing privacy-protecting goals and legislation (Governmental Accounting Office, 2005). Otherwise the GAO has not produced a definitive follow-up report to the 2004 report.

Program Name	Agency	Uses PII	2005 Report	Score
Incident Data Mart	DHS	Y	N	1
Case Management Data Mart	DHS	Y	N	1
Analyst Notebook (I2)	DHS	Y	N	1
Automatic Message Handling System	DHS	N	N	1
Secure Collaborative Operational Prototype / Environment Investigative Data Warehouse	DoJ	Y	N	1
Foreign Terrorist Tracking Task Force Activity	DoJ	Y	Y	2
FBI Intelligence Community Data Mart	DoJ	Y	N	1
OIG – Project Strikeback	DoE	Y	N	1
Insight Smart Discovery	DoD	Y	N	1
Verity K2 Enterprise	DoD	Y	N	1
PATHFINDER	DoD	Y	N	1
Autonomy	DoD	N	N	1
National Cargo Tracking Plan	DoD	N	N	1
BioSense	H&HS	N	N	1
Reveal (*Not introduced in 2004 Report)	IRS	Y	Y	2

"OPEN-NESS" SCORE :
0: Classified Program: few details publicly released
1: Program existence and purpose revealed
2: Data Sources Revealed
3: Scale of Program Revealed
4: Sample Data Records Released (or Sample Screenshots)

Table (1) Operational or Planned Data Mining Programs listed in 2004 GAO Report on Data Mining focused on Counterterrorism with "Open-ness" score and/or 2005 GAO Report.

3. Data Mining Reporting Act of 2007

As part of the "Implementing the Recommendations of the 9/11 Commission Act of 2007," Congress enacted the

Federal Agency Data Mining Reporting Act of 2007[2]. Given the increasing awareness of federal programs and concerns for privacy, the Act requires annual reporting by agencies conducting data mining programs. The Act required reporting on programs meeting the following definition:

> …a program involving pattern-based queries, searches, or other analyses of 1 or more electronic databases, where—
>
> (A) a department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals;
>
> (B) the queries, searches, or other analyses are not subject-based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases; and
>
> (C) the purpose of the queries, searches, or other analyses is not solely—
>
> (i) the detection of fraud, waste, or abuse in a Government agency or program; or
>
> (ii) the security of a Government computer system.

For those programs meeting the definition, among others not included here the following requirements to be reported on were made by the Act [17]:

> - A thorough description of the data mining activity, its goals, and, where appropriate, the target dates for the deployment of the data mining activity.

[2] 42 **U.S.C. Sec. 2000ee-3.**

- A thorough description of the data mining technology that is being used or will be used, including the basis for determining whether a particular pattern or anomaly is indicative of terrorist or criminal activity.

- A thorough description of the data sources that are being or will be used.

- An assessment of the impact or likely impact of the implementation of the data mining activity on the privacy and civil liberties of individuals, including a thorough description of the actions that are being taken or will be taken with regard to the property, privacy, or other rights or privileges of any individual or individuals as a result of the implementation of the data mining activity.

By 2008, the Department of Homeland Security Report on Data Mining provided information on only three programs (Department of Homeland Security, 2008).

4. Lack of Closure on Programs

Since 2004, it has become increasingly difficult to find information pertaining to individual programs named in earlier reports. For instance, of the three programs listed in the 2008 DHS report, none of them can be found in the 2004 GAO report. Similarly, at present, very little information can be found regarding any of the programs listed in the 2004 GAO report (specifically those listed in FIGURE (1)). How these programs have evolved is unknown. The effectiveness of the programs is unknown as well. The reasons for this lack of closure can only be speculated upon. Possible explanations include:

- Less directed oversight by the Congress (as evidenced by the lack of GAO reports published) of data mining programs in general
- Programs have been canceled or defunded
- Programs have become successful and have been classified or reclassified
- Programs have been "rolled up" under other programs and remain "hidden"
- Programs have migrated to new agencies and consolidated with other programs
- Programs have been re-named or consolidated
- National priorities have changed
- Disclosure of program information would jeopardize efficacy of programs

Other reasons may exist for the lack of information available on programs previously known to have existed. Of these, it is plausible that public interest combined with disclosure guidelines have dictated the extent to which agencies release information. Evidence of this exists in the histories presented in previous sections. For example, after 2001, as the government increasingly developed new programs, the public (and Congress) became aware of federal initiatives, and concerns over privacy heightened. In reaction, under Congressional direction the GAO researched all ongoing programs in 2004. Having taken further action with the Data Mining Reporting Act of 2007, Congress narrowed focus on programs affecting privacy and created

guidelines on agency reporting. Any agencies conducting data mining programs not meeting required rules would not need to offer any reports on their activities. Moreover, given public focus on ongoing foreign wars and other matters, awareness of data mining programs and privacy concerns was most likely relegated to the sole attention of privacy organizations.

Whatever the reason(s) for the current lack of information on existing data mining programs, the situation represents a policy problem and presents difficulties to fulfilling the recommendations of the NRC. Without detailed information on the types of data and methods of data mining used in actual programs, it is impossible to fabricate data true to the real data. Attempts to fabricate realistic data would rely on speculation, or perhaps model data mining programs from other fields for which more information is available. Such attempts would not be beneficial to the aim of the NRC report's authors in protecting privacy. Data mining programs under development would not have the synthetic data available for which to conduct testing and research.

D. DHS REPORTS ON DATA MINING PROGRAMS (2008-2009)

A complete survey of federal programs is outside the scope of this thesis. It is germane, however, to highlight some existing programs for which information is available in order to better understand their scope and the data they accept as input. Programs of interest to this study are those utilizing PII and aimed at countering terrorism. The

programs selected are several currently implemented under the purview of the Department of Homeland Security (DHS).

Based on requirements of The Federal Agency Data Mining Reporting Act of 2007, in 2008 the DHS Privacy Office released its report to Congress *Data Mining: Technology and Policy* (2008 Data Mining Report), followed by the 2009 Data Mining Report providing update (Department of Homeland Security, 2008 & 2009). The reports focused on three programs meeting the definition provided in the Data Mining Reporting Act; the Automated Targeting System (ATS); the Data Analysis and Research for Trade Transparency System (DARTTS); and the Freight Assessment System (FAS). The following sections provide more detailed information on the three programs. The ATS will be further broken down to provide more detailed analysis relevant to this study.

1. Automated Targeting System (ATS)

The ATS Inbound, Outbound, and Passenger modules administered by U.S. Customs and Border Protection (CBP), is a comprehensive tool that focuses on travelers, cargo and conveyance methods in order to prevent terrorists or terrorist weapons from entering the United States. As a legacy system (i.e., one that existed prior to 2001 in various forms), ATS also seeks to prevent other criminal activity. The System utilizes information, including PII, about passengers and crew of airliners, vehicles transiting borders, and sea carriers. It also tracks import trends and historical behavior of involved parties. Data is polled, extracted, and aggregated from across numerous databases, public and private.

2. Data Analysis and Research for Trade Transparency System (DARTTS)

The DARTTS, administered by U.S. Immigration and Customs Enforcement (ICE) seeks to detect import/export fraud, money laundering, or other trade-based crimes utilizing anomaly detection to provide leads for further investigation. As a standalone system, DARTTS is not directly connected to any other system but relies on direct input from removable storage media (e.g., CDROM). The data input is comprised of U.S. trade data, foreign trade data, and U.S. financial data. Trade data may be collected from the ATS. Financial data may include PII in the form of individuals carrying in excess of $10,000 while entering the country or making deposits in casinos.

3. The Freight Assessment System (FAS)

The FAS, administered by the Transportation Security Administration (TSA), is a system targeting potentially dangerous cargo aboard passenger aircraft. It uses a rule-based system to identify potentially dangerous cargo and does not use PII. As of the DHS 2009 report, the data mining portion of the system was not yet operational.

These systems represent a small percentage of federal data mining programs, yet each could be considered in itself a major program, comprised of other sub-elements. In order to assess the scope of a "typical" federal program for the purposes of this study, it is necessary to further analyze and review a portion of these sub-elements.

E. FOCUS ON A SPECIFIC PROGRAM: ATS IN DETAIL

A commonality of the three programs detailed in the 2009 DHS report is their consolidated and/or "pyramidal" nature. Each program accesses numerous databases across multiple public and commercial sources. ATS is analyzed further here because it represents a relevant data mining program to the aims of this study, and especially because it is one of the few programs for which more detailed information is available. ATS is comprised of six modules (Department of Homeland Security, 2009):

- ATS Inbound, which tracks inbound cargo to the U.S.;
- ATS Outbound, which tracks outbound cargo from the U.S.;
- ATS Passenger, which deals with travelers aboard various conveyances;
- ATS Land, which deals with private vehicles arriving to the U.S.;
- ATS International, which deals with cargo tracking in cooperation with foreign authorities;
- ATS Trend Analysis, which is an analytical module that processes information from various agencies to detect possible situations that warrant further investigation.

The 2009 DHS report summarizes the ATS in general as follows:

> ATS standardizes names, addresses, conveyance names, and similar data so these data

> elements can be more easily associated with other business data and personal information to form a more complete picture of a traveler, import, or export in context with previous behavior of the parties involved. Traveler, conveyance, and shipment data are processed through ATS and are subject to a real-time, rules-based evaluation.

The report also provides information on three of the six modules per the requirements of the Data Mining Reporting Act. Two of the three modules presented in the report and which use PII are detailed below.

1. ATS Passenger Module (ATS-P)

The Passenger Module of the ATS assesses data on persons arriving or departing from U.S. international ports (to include sea ports). ATS-P is a real-time, rule-based system used by Customs and Border patrol (CBP) officers to identify persons that potentially should not be allowed access in to the country. Data sources include the DHS Advance Passenger Information System (APIS), Non-Immigrant Information System (NIIS), Suspect and Violator Indices (SAVI), Passenger Name Record (PNR) information from commercial airlines, crossing and seizure data from the Treasury Enforcement Communication System (TECS), and the FBI terrorist watch list.

Items of PII collected (in the case of commercial carriers) includes name, date of birth, gender, passport number and country of issuance, passport expiration date, country of residence, travel document type, and U.S. destination address (Department of Homeland Security, 2005). The DHS issued a Privacy Impact Assessment in 2007

proposing new regulations to extend this rule to private aircraft as well (Department of Homeland Security, 2007).

2. ATS Inbound Module (ATS-I)

The Inbound Module of the ATS assesses data on cargo inbound to the U.S. and persons associated with its shipment. The 2009 DHS report summarizes the module as follows (Department of Homeland Security, 2009):

> ATS-Inbound assists CBP officers in identifying inbound cargo shipments that pose a high risk of containing weapons of mass effect, illegal narcotics, or other contraband, and in selecting that cargo for intensive examination…[and]…look at data related to cargo in real time and engage in data mining to provide decision support analysis for targeting of cargo for suspicious activity.

ATS-I does not collect information from individuals but from private entities required by law (e.g., manifests). Data sources include the Automated Manifest System (AMS), the Automated Broker Interface (ABI), the Automated Commercial Environment (ACE), the Food and Drug Administration (FDA), the Automated Commercial System (ACS), and the U.S. Department of Commerce. Items of PII collected include name and address of cargo manufacturer, buyer, seller, and "ship-to" party if different (Department of Homeland Security, 2008).

F. THE NATIONAL RESEARCH COUNCIL RECOMMENDATION

In a study of data mining for counterterrorism and government programs in general, the National Research Council of the National Academies produced a comprehensive report on the subject of data mining for counterterrorism

and privacy (National Research Council, 2008). The previously outlined events leading up to the publication of the NRC report in 2008 contributed to a public environment and political culture concerned with individual privacy. The title of the report, "Protecting Individual Privacy in the Struggle Against Terrorists" is illustrative. The following recommendation was included in the report:

> To protect the privacy of innocent people, the research and development of any information-based counterterrorism program should be conducted with synthetic population data. If and when a program meets the criteria for deployment in the committee's illustrative framework described in Chapter 2, it should be deployed only in a carefully phased manner, e.g., being field tested and evaluated at a modest number of sites before being scaled up for general use. At all stages of a phased deployment, data about individuals should be rigorously subjected to the full safeguards of the framework.

No discussion of how this data would be created or its uses is offered in the report. The NRC also recommended strict audit and oversight of any such programs using PII.

G. CONCLUSION

Use of data involving PII may result in situations where innocent individuals are targeted for surveillance or investigation based on poor conclusions a data mining program may have made. Such false positive conclusions may be made as a result of an individual's records or behavior matching preconceived rules of a system but which are otherwise not criminal in nature. The NRC report maintains:

> …even in well-managed programs such tools are likely to return significant rates of false positives, especially if the tools are highly automated. Because the data being analyzed are primarily about ordinary, law-abiding citizens and businesses, false positives can result in invasion of their privacy (National Research Council, 2008).

Individuals falsely identified as criminal, or even as possible suspects considered otherwise innocent, may then be subjected to unwanted surveillance (Garfinkel, 2006), public ridicule, detention or worse. The GAO offers an example in its 2005 report on Data Mining of such situations. Lists of names of suspected "suicide bombers" were given to the FBI by foreign governments. The FBI subsequently investigated the lists to ascertain if any of the names matched or were similar to names or persons residing in the United States (Department of Homeland Security, 2005). Such situations could lead directly to the surveillance of innocent persons. This scenario also presents a further nuance of the challenges of privacy and data mining. A suicide bomber only carries that description after they have committed the crime. Identification of persons who might commit such an act identifies what remains an otherwise innocent person.

The Automated Targeting System of the Customs and Border Protection agency may be considered representative of a typical federal data mining program. As a subject of study in the scope of this thesis, it presents an illustration of the great scope of the challenge of fabricating data for the purpose of developing new data mining programs per the National Research Council

recommendations. It is noted, however, that the ATS is a Program built upon other programs and developed over a period of many years, refined across emerging technologies, and consolidated across various agencies. As such, it may not be representative of the singular data mining program that a federal agency might seek to implement or add to its arsenal. An attempt to create synthesized data on a smaller scale for a singular program might pose a more realistic challenge as a foundation for this research.

IV. DATA MINING RESEARCH AND SYNTHETIC DATA

A. PURPOSE OF SYNTHETIC DATA

Given the extensive use of data mining by the government for detection of terrorist activity, it is reasonable to expect that current programs may be expanded or modified, and that other programs will be developed in the future. Per the NRC recommendation, such programs would benefit from the availability of synthetic data protecting the real private data of persons.

How will these programs be tested? How will they be certified (and using what data) by other or outside agencies, if required? Will existing databases containing the PII of average citizens be used as input to test the programs? For example, suppose a government agency developed a new data mining program that accepted as input the names, addresses, tax information, and prior gun purchases of individuals. Would that agency have a reasonable expectation that other agencies such as the IRS and FBI readily provide them with existing databases from which to test their new program?

Some might argue that the fusion of such data in itself creates a scenario in which individual privacy rights are violated. Even in a testing/research environment, in this scenario a new data record would exist that locates an individual, points to their earnings and tax forfeitures, and lists their firearms possibly located in the home. Such a record would presumably not only violate a reasonable person's expectation of privacy, but

would also violate existing privacy laws[3] that regulate the government's access to non-governmental data.

Data containing PII does not solely exist in the public sector. While data such as individual driver's license data and tax records are stored by the government, other data, such as credit history or details of credit transactions, are held in the private sector. Private data is already used as input to federal data mining programs (Governmental Accounting Office, 2004). Regarding sources of data in general, the National Research Council (2008) concludes:

> The government collects information from many industry and government organizations, including telecommunications, electricity, transportation and shipping, law enforcement, customs agents, chemical and biological industries, finance, banking, and air transportation.

When combined, data from focused mining of such data stores may provide clues pertaining to a broader picture of an individual's activities. Extending the previous example, in addition to tax records and gun purchases, the same program might track hardware store purchases to find large purchases of fertilizer. An average citizen might not object to the government learning they purchased a new hammer, but a farmer buying fertilizers might object to appearing on a list of potential terrorists, along with his yearly earnings and previous gun purchases.

[3] The Brady Handgun Violence Prevention Act of 1993 (Public Law 103-159) mandates destruction of handgun sales transactions by federal agencies after the approval/background check process is completed.

Using – or even collecting - data in this manner, however, may violate the individual's right to privacy. While the pursuit and detection of terrorist activity has been identified as an important part of protecting citizens from calamity, a very fine balance between law enforcement activity and preservation of individual privacy presents itself as a primary goal of a democratic society.

Given the need identified for terrorism detection, combined with the potential usefulness of data mining across a vast spectrum of available data, data mining may present a useful tool. The use of data involving personal information of U.S. citizens however remains subject to the fine balance between the need for privacy and the desire for detection of malicious activity. If it was possible to create a useful and realistic body of data representative of actual PII, it could be used to develop data mining programs without concern for violating privacy rights of individuals.

B. DATA CHARCTERISTICS

To be valid for use in data mining program evaluation or research, synthetic data needs to be both realistic and useful. A third category that is related to realism and usefulness is fidelity.

1. Realism

To be realistic, data needs to statistically mirror real population data. For example, if considering a database containing airline passenger information, individuals on a flight manifest would ideally be broken

down proportionally into realistic percentages of country of origin and ethnicity based on real flight data. Other attributes important to realism in this example might include age, gender, and final destination. Flight manifests would need to mirror accurate numbers of passengers per real flight data. To create such realistic data, extensive statistical data based on real data would need to be attained and/or calculated prior to creating the model used to synthesize the data.

An added complication is the interconnection between multiple data sets. Effective data mining does not evaluate data in isolation. In this example, most likely other data would be correlated against the flight data. To be realistic, the flight data would need to be matched by other synthetic databases. Such databases could include banking or immigration records, for example. Secondary databases such as these would need to include some of the same synthetic individuals as the others that are present in the primary dataset. And, as with flight manifests, any secondary datasets would need to have some metric to measure realism.

Interestingly, any statistics based on real data could present similar problems with "anonymization" as mentioned in previous chapters depending on the depth of the data. Herein lays a potential paradox: the more that synthetic data relies upon real statistics, the greater the chance that mirroring real data may leak information about the real datasets on which it is based. At one extreme lies the potential of re-identification. At a lower extreme,

other information could be leaked, such as the number of children in a condominium association, or the minimum and maximum income of a person living in an apartment building. Given the primary goal of creating synthetic data to prevent violations of privacy, any attempt to create such data would need to be cautious in this regard.

2. Usefulness

Realism alone is not sufficient. Synthetic data for evaluation would need to be comprised of a statistically significant number of entries. For example, a synthetic driver's license dataset containing 100 entries used to test a data mining program at the state population level would not be statistically significant.

To be useful, data would also need to be presented in a certain form. Perhaps to be realistic the data would be created in such a manner as not to be initially useful to the program, and require modification to faithfully represent reality. In this case, research into the format of data output by actual systems intended to be used as input into the new program would need to be thoroughly understood and copied; or perhaps the data would need to be created in a manner consistent with the input format of the program. Such decisions would be dependent on the goal of the data mining program under development.

C. DATA SYNTHESIS; APPROACHES TO MODELING

Creating fake data involves creating some model in which to simulate real data. Models could take many

approaches and utilize various technologies. Three possible approaches to modeling synthesized data are presented below.

1. Manual Creation

In this approach, a human operator simply creates data in a "brute-force" manner according to the requirements of the data mining program. For example, the model might call for creation of a 30-year-old man who resides at some address. The creator would simply manufacturer the record creatively.

2. Automated Direct Creation of Realistic Data

The Automated Direct Creation of Realistic Data model is one in which the model outputs only the data that will be used in an arbitrary data mining research program. In other words, the model does nothing more than synthesize and concatenate strings ideally representing some meaningful approximation of realistic data.

For example, such a model drawing from some resource of meaningful names or rule-based algorithm might produce a record:

<FName, LName, Street Address, City, State>

for a non-real individual. Such records would be generated by the model and represent its only output. Every pertinent event to an individual data item history would be created explicitly. As in the previous example, this model would also create some 30-year-old man, however some automation would be involved. Rather than a human simply

inventing the person, the data forming the elements (e.g., name, address, etc.) could come from some other existing database(s), or be randomly generated within certain parameters.

3. Generative Simulation

In contrast to the previous models, the Generative Simulation model would need to apply a methodology taking into account the notional world in which the data "resides" and a simulation of the passage of time.

Utilizing the previous example, in the simulated "world" in which the data resides the fabricated individual is 30 years old. But what happened to that individual 30 years prior to the present? With the manual model and Automated Direct Creation of Realistic Data model, a 30-year-old individual would be created. With the generative model, the individual might be born 30 years in the past to two simulated parents and allowed to "grow" within the simulation. The person would presumably have an address, school, and other environmental variables potentially affecting his/her attributes at age 30 after 30 years of simulation.

The generative model not only produces the desired datasets for input into the hypothetical data mining program, but also contains hidden variables and rules used in the conduct of the simulation and which are not reported in the outputs. The simulation is based on some objective "ground truth" world model that produces the resultant

data. These values could be used to validate the outputs of the data mining program, or to suggest new research directions.

Urban Sim employs such a generative model. Urban Sim allows city planners to take an existing city model, apply changes in the present (e.g., build a bridge), and simulate the impact the changes may have on the city in the future (Waddell, 2002). Urban Sim is not a model conducive to synthesizing data for counterterrorism, since it does not generate data at the person level.

4. Comparison of Approaches

The models presented above are categories of models. They could be implemented in many different ways. As broad categories however, the two represent a relatively complete grouping of models that could be used to synthesize data. This section further explores the models at this higher level of abstraction without discussing technical implementation details.

In terms of realism, the merits these models are heavily dependent on the merits of the approach used in their development. In other words, a good implementation of one would always be better than a bad implementation of the other. In this way, the models have several similarities. Each would be subject to the bias (or lack of bias) of the creator(s) and built-in rules of the simulation. They would require careful attention and decision to types of names, gender, ethnicities and creeds produced in the output, depending on the goals of the

program. The models would also require decision as to the sharing and protection of datasets based on the level of realism of the data. Finally, both models would be required to produce data of sufficient depth and texture in order to be useful to a data mining program.

An argument supporting the merits of the Generative Model comes from the field of Computer Graphics. Creating simulations of moving clouds, humans, or other subjects encountered in reality, researchers found success in injecting underlying data variables also based on reality. For example, a simple simulation of a human walking may not initially look realistic to a real human observer. After providing the simulated human with data representing bones and veins, however, the walking simulation gains an additional measure of realistic movement. Albrecht, Haber, & Seidel (2003) provide an example of this type of simulation, modeling human hands using an underlying anatomical structure. In order to realistically simulate hand movements, models are provided with realistic data simulating such variables as skin elasticity and muscle contraction.

An argument against the use of the Generative Model is the difficulty in effecting its implementation. If not properly executed, a generative model might produce unrealistic data over time (i.e., simulated time). For example, a poorly implemented model might create an individual attending University for 20 years or who has multiple spouses. The quality of the resulting output, used as input to a data mining program in development,

would be non-deterministic to a great extent. Assuming a large volume of data was produced, evaluating the quality of the data would be difficult. The actual evaluation of the quality of the data might not be undertaken until actually used as input to the data mining program. If the data was faulty, instead of testing the abilities of the program under development, the program itself might apply or waste its resources simply on finding the errors resident in the test data set.

D. DATA SYNTHESIS; ADDITIONAL CONSIDERATIONS

1. Creating Names

Given the desired task of creating data mirroring PII in a realistic way, a primary data item would be names of people. Depending on the methods used to synthesize names, the potential exists to create names that actually exist in reality. If synthesized names match real people's names, it may present privacy concerns or create public relations problems for the sponsoring agency. This may or may not be an issue depending on the aims of the research/program.

One method of creating names might be to create separate databases containing last names and first names. Sources for the databases could be realistic sources such as telephone directories or non-realistic sources, such as a baby name book sold to expectant Parents for first names and lists of common last names (e.g., data sources listing surnames by geographical area for genealogists). The model would concatenate data items from each of the two databases to create simulated people having realistic names, first

and last. In either case, it would be possible to create the name of a real person, though for sake of argument using a telephone directory would most likely have a greater probability of creating a match to an actual person.

A second consideration is the characteristics of the names created. To avoid matching genuine (i.e., real) names, a model might produce names comprised of gibberish. For example, the unlikely and ethnically ambiguous name "Wicoiu E. Gjdoi" could be generated and used. A problem with this approach is that, to human analysts developing or testing a data mining program, such names contain no extrinsic information such as ethnicity, national origin, etc. Such information might be relevant to a data mining application. Without any historical or ethnic basis, all of the names created would generally be indistinguishable from each other. Depending on the goals of the model, such an approach might not be helpful.

To be realistic, a model might attempt to create names that mirror ethnic reality. The model might have many ethnicities represented, with first and last names deliberately paired together. For example, a realistic model might not produce the resultant name "Seamus O'Muhammed" but instead "Seamus O'Toole." As previously mentioned, the more realistic the names become the greater the chance would be of mimicking real persons. Using ethnicities also might create the impression of ethnic targeting/profiling, depending on the model.

2. Protecting Data

Thus far, discussion of which organizations, agencies, or institutions might create synthetic data and for which purpose (research or operational use) for counterterrorism has not been explored. However, it could be presumed that such a body of data would be created for the purposes of research. Assuming a useful body of fabricated data was created, considerations might be needed to be taken in terms of protecting the data.

Depending on the level of reality of the data (as per the previous two paragraphs), some overlapping of reality (such as real names of real people, or real social security numbers existing in the data) might occur. In such a case, open release of such a body of data might not be appropriate.

A second problem with releasing a body of data created for counterterrorism could lay in the nature and revelations of the data. Depending on the model(s) used to generate the data, releasing the data could present several problems. First, if the data was geared toward a certain data mining program, the details or mechanisms of the program itself might be revealed in the data. Such revelations could compromise the investigative strategy behind the program. Second, data that was infused with deliberate terrorist targets could be viewed by civil libertarians of indicating a political or racial bias (presumed or otherwise) of the program and/or its creators.

V. DATA MODALITIES

The broad term *data* refers to numerous categories of information. From distinct groups of binary digits to shades of black and white in a traditional photograph, or from noise energy to neuro-responses felt on the skin, the term data can mean many things. In the lesser scope of computer science, counterintelligence, and data mining, the term still comprises a large range of varying types of information. It is, therefore, useful to discuss the types of data that might be synthesized for data mining purposes.

A. TYPES OF DATA

The following represent generally broad categories of data that are available or may be used for the purposes of counterterror.

1. Record-Based

Record-based data represents the traditional "database" collection of records increasingly used throughout the information world (Fayyad & Uthurusamy, 2002). Records include anything from names and addresses to amounts of money in an account, height and weight, or social security number. Records are generally anything created at a point in time that, once recorded, remain useful as representative of fact. For example, if an individual's social security number is established and recorded (as record), even if it is subsequently changed, the initial record remains a fact established. The fact may or may not be useful. Data representing records

generally take on an accepted form useful to the using entity (e.g., manually recorded by implement in established fields on a paper form, or ASCII (digital) representation readable by a human on an output media).

Records are useful to data mining in several areas. First, when stored in digital media, records represent a simple way to effect the process of data mining based on the characteristics of record-based data. That is, such data can be represented in very few characters (e.g., numbers and letters) which are generally packaged in distinct and identifiable patterns. As such, algorithms can be created to manipulate these finite patterns in meaningful ways and in useable time.

Record-based data represents perhaps the most accessible format for synthesis and for use in a data mining program. It offers a typical genre of information used to define people, places and things, and is a mainstay of most databases having specific defined fields. Given the predefined characteristics of record-based data, as well as its composition of a limited character set, record-based data provides a useful format for synthesis. Names of people, places of residence or birth, or identifying numbers each represent finite items that can be mimicked. One aim of this thesis will be the exploration of the practicality of synthesizing record-based data. A brief discussion and speculation of the feasibility of synthesizing other types of data is included in the following sections.

2. Photographic Data

Photographs, whether from traditional technology using photographic paper or modern digital technology, can be categorized as data able to be interpreted by the unaided human eye. That is, photographic data (when properly displayed) mimics data the human eye views in nature and is thus meaningful to the human brain. Digital photographic data has the added characteristic of its being represented (by definition) in binary (or other) form.

Photographic data is useful in counterintelligence as a record of something that occurred or a representation at a specific point in time of the state of what was recorded in the photograph. Such benefits may include geo-location, temporal location, or the intrinsic value of the photographed object itself. It may identify an individual of interest from among a larger body of data. As such, it is useful to data mining technology for counterterrorism. Using facial recognition features, researchers have demonstrated the ability to match the identities of individuals from photographic data (Zhao, Chellappa, Phillips, & Rosenfeld, 2003).

Synthesis of photographs may present an effort without reward. While possible to the extent that utilizing computer generated graphics technology may generate illustrations mimicking photographs, the time and creative energy involved, coupled with the limited output, may not prove valuable to a data mining approach.

An alternative to generating synthetic photographic data lays in the potential use of actual photographs. For

instance, photographs of real, deceased people, such as are available in Civil War photos, could be scanned and digitized to provide a useful body of data for a program. Depending on the requirements of the research, representations of such persons could be inserted into other photographs or backgrounds having certain locations, people or things.

3. Biometric Data

Biometric data represents the unique features of an individual -- a living entity. It relies upon modern technology (e.g., retinal scanners, fingerprint ink, etc.) to record these characteristics unique to the individual. Data, such as individual DNA or dental records, may be used to distinguish one individual from another.

Similar to photographic data, biometric data is useful to data mining insofar as a person of interest may be located or identified from within a larger collection of data. Like photographs, biometric data may be represented in digital form. As such, data mining technology may be applied to the data.

Biometric data could take on various forms. While a fingerprint or retina could be viewed as an image, for coherence in a data mining program, each unique image could be represented as some sequence of numbers. As such, a collection of unique numbers could be generated. Whether or not they would be useful is beyond the scope of this thesis.

4. Narrative Data

Narrative data is prose recording events or stories known to a person (whether fact or fabricated). Narrative data might be written or recorded on some voice recording media. The data may contain useful information regarding events or individuals. Examples include police reports and eyewitness accounts of events. Narrative data is typically not created by the would-be subject of a data mining exercise.

Narrative data is useful to data mining for counterterrorism. Whether contained in voice recorded media or written (and transcribed to digital media) the data can be searched for items of interest such as names, person descriptions, vehicle descriptions, modus operandi, and so on. The University of Arizona COPLINK project demonstrated some success using a neural network–based entity extractor for data items (such as narcotic or criminal names) from police narrative reports (Chen, et al., 2003).

Narrative data could be created primarily using a creative approach. Like an author writing fiction, a human creator of narrative data would manufacture the information. This approach would be inefficient and perhaps not useful to a data mining program, unless a small amount of data were required. Any other machine-based methodology would either require a model that could approach the capabilities of a human or deliberately embed information in existing writings, such as novels or articles. Existing narratives could be repurposed and/or

anonymized. Whether the latter would be useful or mirror genuine narratives would require further research. The original ELIZA project and subsequent work remains an example of the difficulty of machines in mimicking humans (Weizenbaum, 1983). Further, "chatterbots" such as ELIZA also rely on input from a human user in order to maintain some amount of realism.

5. Free Format Data

Free format data is not limited by fields or format rules. Examples include emails, Web pages and voice recordings. While similar to narrative data, it is distinguished here as a separate category in terms of its characteristics. While narrative data generally tells a story, having a subject and timeline of events, free format data can include any non-formatted or fielded character.

Free format data could prove very useful to the task of data mining for counterterrorism. Assuming a useful body of synthesized e-mails could be produced, the data simulating communications within a malicious social network could be used as input to a program. But like photographs, the effort to create free format data might not yield sufficient reward. Genuine free format data could be used that does not contain PII or is not governed by privacy laws. Additionally, some amount of data already exists for use in research, such as the ENRON emails. Such existing data might be useful itself, or could be modified for use in a specific program.

Like narrative data, generating free format data would require some creative approach, presumably by a human, unless some body of data containing non-meaningful information was desired.

B. SIMPLIFIED TEST DATA

A miscellaneous category of data is data specifically created for model testing purposes. It can be differentiated from other synthetic data based on scale and complexity.

Simplified test data might be required to test certain functional areas of a data mining program in development. For example, if testing the ability of a program to read names from a file and redirect them to a database in usable form, the test data might contain certain characters more likely to create errors in the program code, such as an underscore or percent character, accented characters, and so forth. Otherwise, the data might not have or need any realism in terms of fidelity to actual names of persons.

Test data might also be synthesized for more complex scenarios. Data may be required to test a program's ability to extract or identify specific elements. For example, datasets created with known targets hidden in the data could test the efficacy of the program, to perform phonetic name matches, to handle typos, and to winnow out deliberate near-matches that might produce false positives that may also be deliberately embedded in the data.

Test data might also be used to test the ability of a program to detect anomalies in data. For example, a data

mining program might seek to detect variations in names. A targeted data item such as the last name “Jones” might be resident in data as “janes” or “jonas.” Test data could be created to mimic such errors that might be encountered in reality.

VI. A DATA SYNTHESIS EXERCISE

A. REPRODUCING THE DATA

The descriptions provided in Chapter III regarding the Automated Tracking System (ATS) represent a limited but illustrative view of the extensive scope of some federal data mining Programs. As with the ATS, a new data mining computer program would feasibly source information from numerous databases. These databases are in practice often updated in real time. As with the ATS, which sources PNR from flight data, it can be presumed that such updates occur constantly and daily. The source data would most likely contain numerous fields and a vast number of populated entries. Moreover, such a program might contain several technologies performing across several commodity areas, as with the ATS.

Synthesizing data to test a new system such as the ATS would require an incredible effort of research and production. In addition to populating data consisting of fabricated PII across the numerous databases used by the program, the data itself would need to meet certain requirements; specifically, the data would need to be created in such a way that it was realistic and useful.

1. Deconstructing the ATS

In order to further explore the possibility of creating synthetic data for counterterrorism data mining Programs, an exercise based on the entire ATS would be prohibitive. Using the ATS as a genuine example of federal

data mining programs does offer a basis for an exercise on a smaller scale however. Deconstructing the ATS offers some insight into its subcomponents, although available information is limited. Taken together, the DHS Office of the Inspector General (OIG) report "Review of the Immigration and Customs Enforcement's Compliance and Enforcement Unit" (2005) and excerpts from the Federal Register (2006)[4] provide some information on the extent of the ATS and related systems.

The Federal Register provides information on the ATS, while the DHS report provides information on three programs specifically related to border protection, including the National Security Entry Exit Registration System (NSEERS), the Student and Exchange Visitor Information System (SEVIS), the United States Visitor and Immigrant Status Indicator Technology (US-VISIT), and a fourth "environment" termed the Treasury Enforcement Communication System (TECS).

Interestingly, the Federal Register (reporting solely on ATS) does not mention any of the three border protection systems, and the DHS OIG report does not mention the ATS. Yet both mention the TECS, which appears to be a predecessor to the ATS, though ATS has not replaced TECS. Presumably, some migration of terms and systems may be in evidence here related to the creation of the DHS and resulting agency consolidations. Regardless, it is realistic that the three border protection programs are

[4] Federal Register: November 2, 2006 (Volume 71, Number 212).

related to, if not a part of, the ATS. The Federal Register states:

> ATS involves the collection and creation of information that is maintained in a system of records. Previously, this information was covered by the…(TECS) system of records notice, as ATS is a functional module associated with the environment of TECS. ATS is employed as an analytical tool to enhance CBP screening and targeting capabilities by permitting query-based comparisons of different data modules associated with the TECS system, as well as comparisons with data sets from sources outside of TECS.

Further, Appendix (A) of the DHS OIG report (2005) offers some illustration of these systems and their relation to investigatory workflow.

2. A Realistic Data Group to Populate

"Drilling Down" further on one of the three border protection programs, the SEVIS program provides yet a narrower example on which to focus this study. SEVIS represents the "technical part" of the Student Exchange Visitor Program (SEVP). SEVP provides data on foreign students obtaining non-immigrant visas in order to study in the United States. Through SEVIS, it further allows the government to track authorized dates (and deadlines) of the foreign student "stays" within the U.S. On a broader scale, SEVP includes partnering Universities, which carry some responsibility to administer SEVIS and must provide trained individuals to implement the program. No foreign student may study in the U.S., except under SEVP and such partner universities.

A main input to SEVIS is a form potential students must complete prior to their entry into the U.S. in order to pay an applicable SEVIS fee. The DHS I-901 "Fee Remittance for Certain F, J, and M Nonimmigrants" form collects several items of PII from the individual (see Appendix A). The data on this form presents a realistic (actual) group of data fields useful to a data synthesis exercise; it represents a known input to possible federal data mining programs.

3. Available Statistics to Support Realism

In order to attempt to mirror realism in the chosen body of data, real statistics governing student populations are helpful. Presumably, gathering actual statistics regarding the entire foreign student population in the United States is not possible (except from the SEVIS databases themselves). Continuing to narrow the focus of this study in the interest of practicality, the student population statistics of one university would be instructive. The California State University, Long Beach (CSULB) provides statistics on foreign students. Appendix (B) provides a sample of the statistics available; other statistics including gender, ethnicity by degree level, citizenship status, and other such tables are also provided.

4. Secondary Database

To mimic a real data mining program, some secondary database is required to accompany the student database for input to the program. Candidates could include flight

manifests, local phone books (with addresses) or campus directories, or driver's license information, among others. Given the likeliness of a foreign student desiring to operate a motor vehicle while residing in the United States, the California Driver's License form was chosen as a secondary information source. Appendix C contains a sample application form.

This exercise has attempted to achieve some degree of realism by utilizing details of a known data mining program, namely the ATS. No available public information has been found that indicates whether or not ATS has access to (or uses) driver's license data as one of its input databases. For the purposes of this exercise, however, it is reasonable to expect that data such as that included in a driver's license record, if not actually used in a program, is closely related to genuine inputs. Moreover, one of the primary aims of the exercise is to ascertain if synthesizing data is possible and useful. Using driver's license information as a secondary data source for the ATS provides an avenue to achieve this aim.

B. MODEL SPECIFICATION

The following model represents one possible and simplified approach to synthesizing data. The model attempts to capture a small portion of the ATS as outlined in previous sections; it does not attempt to synthesize input data for the entire program. Lastly, it is not presented as a usable data generator (of which many already

exist)[5] but as an opportunity to explore methods of data generation for comparison and illustration for the purposes of this thesis.

As mentioned previously, many details of the ATS program are unavailable. Additionally, using two simple record-based data sources offers a realistic starting point to assess feasibility of data synthesis. The chosen data sources are the SEVIS fee form for foreign students (Appendix A) and the California driver's license application (Appendix C). For purposes of brevity and practicality, not all fields of the chosen data sources will be applied. A diagram of the model is seen in FIGURE (2). The central focus of the model is the abstract data item representing an individual person.

For purposes of brevity and practicality, not all fields of the chosen data sources will be applied. A diagram of the model is seen in FIGURE (2). The central focus of the model is the abstract data item representing an individual person.

1. Inputs

The foundation of the model is based upon "seed" records. The seed records are the most raw input of the model (i.e., the "building blocks" for generating more comprehensive data) used to generate full records representing individual persons. Seed records are characterized by having a minimum number of characters to represent one full field item and other associated values.

[5] For example, Turbo Data available at: www.turbodata.ca.

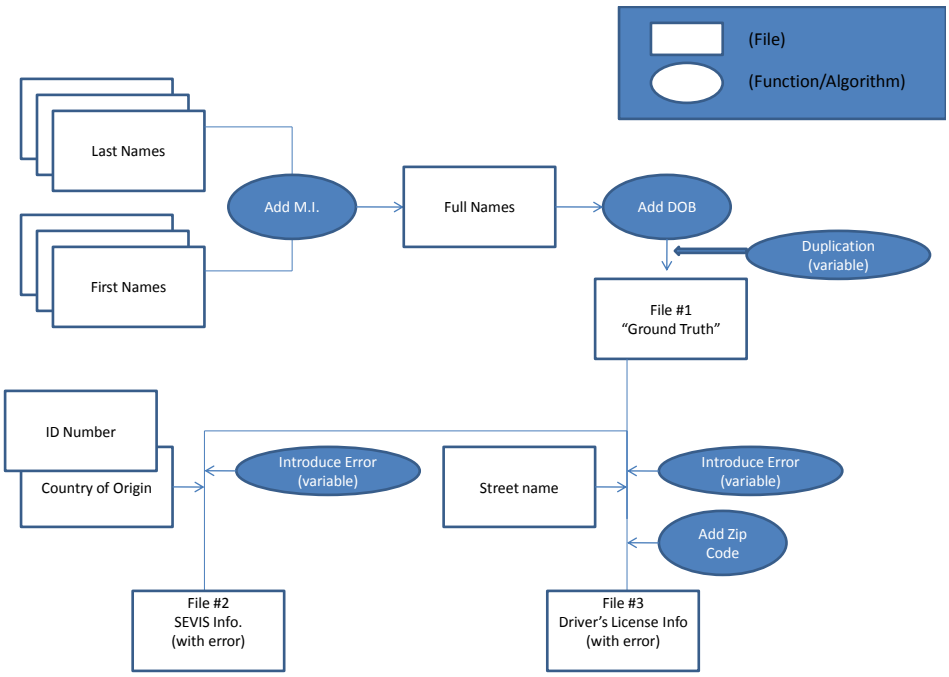

Figure (1) Model Overview, Automated Direct Creation.

For example, an individual field in a seed record would be one last name. Associated values would include items such as nationality and country of origin. The purpose of the seed records is to provide the model with base data from which to generate full data representing persons. Seed records are contained in files grouped by category such as "last names." A surname seed file (.csv) might look like that contained in FIGURE (3), where names are assigned unique identifiers for ethnicity, such as 'E' for English, and country of origin, such as 'G' for Germany.

```
LName, Ethnicity, CountryOfOrigin
Jones, English, US
O'Connor, Irish, UK
Deutsch, German, DE
Nowak, Polish, US
…
```

Figure (2) Sample "seed" surname file.csv

A similar seed file for first names would provide the input from which to match with surnames.

The sources of the seed record files vary. In order to minimize creation of genuine items of PII, files are either manually generated or sourced from on-line records (such as available lists of baby names for expectant parents). Further, no seed records are sourced from data containing more than one seed item. That is, first names attributable to last names, or vice versa, from any source are not used in order to avoid reproduction (or even the perceived reproduction) of a real person from a valid data source. For example, a record containing the string "Jim Strom" is not used. The source of the record could ultimately hail from some genuine data source traceable to a real person. Only records containing first names or last names are used.

Numbers, including identification numbers, street numbers, ages, etc., are generated pseudo-randomly where appropriate, and within parameters based on applicable rules. For example, the age of an individual attending a university would be generated based on student population

statistics (Appendix B) and fall within rules, such as the age not falling below 17. Otherwise, the program would assign peoples' ages within these parameters.

2. Outputs

The output of the model includes three main groups of synthesized data records presented in long form and collected in three files. The first group represents the "ground truth" data. This data provides what is true and accurate with regards to the data in the simulated world. It provides a starting point to generate the two other files and introduce error. The below illustrates the format of the three groups:

(1) <LName, FName, MI, DOB, Nationality>

(2) <LName, FName, MI, Nationality, DOB, SEVIS_ID#>

(3) <LName, FName, MI, Street#, StreetName, ZipCode, DOB>

The output files demonstrate realism based on a mix of ethnic names matching statistics as provided in Appendix (B). The length of each output file is dependent on the length/depth of the inputs.

3. Treatment of Inputs/Record Creation

Name (first and last) seed files, delineated by ethnicity, are concatenated randomly to create full names using the Person Class defined in FIGURE (4). Middle initials are assigned randomly, to include some records having no middle initial. Names are created within rules governing ethnicity. For example, seed files containing

Polish surnames would be concatenated only with Polish first names, where applicable.

Person Class

FirstName	(String)
LastName	(String)
DOB	(Date)
MI	(String)
Ethnicity	(String)
Nationality	(String)

Figure (3) Class:person

Using the person class, a file of persons is generated from the seed files. For instance,

Create_a_person(nationality, age)

set first name from names table with nationality
set last name from names table with nationality
generate middle or no initial
set DOB = year of simulation – age
set country of origin

The nationality variable allows the simulation to reflect the weights of nationalities in a group of people via a configuration file in order to mirror statistics representing reality. For example, using Appendix (B) as

input, weights (from 2009) would list 2.5% Australian, .4% Brazilian, and so forth. A list is created mirroring the percentages. For example,

List of nationalities = []
    For each nationality in configuration file:
        Add nationality to list 100(%) times*

Given a list populated with accurate percentages of nationalities, output files are generated reflecting the statistics inherent in the tables/statistics and with some variation:

Weighted_nationality:
    N = Random draw of nationality from list
Weighted last_name(N):
    Random draw of last name for nationality = N

Across a statistically significant number of persons, the weights of nationalities reflected in the output file would more closely resemble the statistical base.

The age variable is similarly passed through the create_a_person method. Given preset parameters such as,

Student_min_age = 18
Student_max_age = 30

The age is chosen and Date of birth generated. For example,

```
Get_DOB:
     Month = random month
     Day = random day
     Year = current year –
          (random[student_min_age:student_max_age])
     DOB = Month + Day + Year
```

Or instead of choosing a random year for the age, statistics could be applied based on real data.

4. Deliberate Introduction of Error

Some level of error within data records is desirable. As per reality, error in data might be introduced at the point of administrative data entry or by intentional action of an individual. Data mining programs must negotiate errors and variations in data.

The amount of error introduced in data synthesis is controlled by an adjustable "error" parameter based upon the needs of the program. For example, a selection of [.05] would present a 5% error rate within selected fields of data. The error is introduced to the data file in the creation of a number of persons. For instance:

```
Error rate = .05
Make 1000 people:
     P <- create a person(weighted nationality, age)
     If random < Error rate, make_error(P)
     Append P to output file
```

To maintain record of the fact that an error is an error amongst a large body of otherwise created data, the correct version of the data upon which the error is based must be extant and recallable. The use of the ground truth file (see FIGURE 2) allows for the maintenance of correct data. Errors may be introduced in either or both files from the ground truth file. Once full name records are generated, a modification may be applied to a field in the record and passed to the resulting data files with error.

5. Control of Duplication

Some duplication of data within synthesized records is desirable. For example, two students might have the same address, given that they share a residence. Deliberate duplication in such cases is included in the model in a controlled manner. Duplication may be inherent depending on the technique used for creating names. Some duplication of last names is also desirable, as seen in reality.

A design choice is whether duplication is randomly generated or targeted. Targeted duplication is focused on various data groups (e.g., certain last name ethnic groups), where certain last names are deliberately copied to create a new individual. Random duplication is simply based on a percentage variable applied across all data. For example, a choice of (.02) would choose 2% of names and copy them to create new people. For instance, as in introducing error,

```
Duplication rate = .02
Make 1000 persons
     P <- create_a_person(weighted nationality, age)
     If random < Duplication rate make_dupe(P)
     Append P to output file
```

Undesirable duplication includes ID numbers, which must be unique. Simple incrementing of ID number creation parameters within the model ensure unique data per individual.

If real data is available, error rates matching that seen in real systems could be used to generate the model. For testing purposes, error rates could be deliberately elevated in order to test the "coping" mechanisms of the program under development.

6. Limitations of This Model

This model presents perhaps the most simplistic approach to synthesizing data within the framework of data mining for counterterrorism. As such, it is useful as a baseline model.

The model is limited in several areas. The model only contains data from two original sources. Further, the data from the two sources is limited in terms of number of fields presented (whereas the original source documents have more fields available). The outputs are also limited based on quantity and quality. The quantity is limited by the number of seed files which must be manually created. The quality is limited to the creativity and completeness

of the input seed data. For example, the statistics presented in Appendix B demonstrate many different nationalities. Whether or not all of these nationalities are recreated in the model and assigned realistic ethnic names (requiring further research) affects the quality of the outputs.

Street names and numbers are overly simplified. Following the approach of the Urban Sim model, a genuine city grid could be used to represent actual streets and numbers. The quality of the data would depend on the level of detail of the model. For example, is it important if a certain "target" student lives at an address next to, or within sight of, another student? Is it important to the program whether or not a student lives within walking distance of the University or some social gathering site? If desirable, actual data could be taken from a Geographical Information System (GIS).

THIS PAGE INTENTIONALLY LEFT BLANK

VII. FURTHER CONSIDERATIONS AND FUTURE WORK

A. CHALLENGES TO CREATING SYNTHETIC DATA

Previous chapters have outlined many of the challenges and issues inherent in creating synthetic data for use in data mining. A concerted effort undertaken to create a useful and realistic body of data involving PII may encounter other difficulties or potential barriers.

1. Effort of Human Designer

Developing a model to explore possible methods of data creation (such as that presented in Chapter VI) offers some insight into the difficulties of creating synthetic data. One major difficulty may arise based on the scale of the data required to test a major data mining program under development.

The approach outlined in Chapter VI requires considerable effort upfront (i.e., prior to automation) on the part of a human actor/designer to create (or source) the input files. The input files must contain a number of first and last names of a quantity directly related to the number of full names desired in the output. Moreover, they must be correlated by ethnicity. The human designer is required to understand which names correlate with which ethnicity. Assuming the designer has automated a process to input presorted files of ethic names by group into the model, these files still must be located (if available from a third party/source), classified, and possibly manipulated to make usable.

Assuming that a realistic and useful body of synthetic data would be very large, mirroring to some degree the number of records available to a data mining program in reality, the amount of effort required by the designer could be prohibitive. At some level, the human effort required to create a large number of names might negate the usefulness of this approach (automated direct creation), making it similar to the manual method in terms of time and effort.

2. Perceived Privacy of Data

Another possible issue resulting from the vastness of scale arises with regard to perceived privacy. The greater the amount of "fake" people created, the greater the chance of creating individuals who mirror actual people, to include full names, dates of birth, etc. To illustrate, assume a body of data is created to meet requirements of some program for output of one million names/records. If, for example, 2% or 20,000 records are created that exactly match the PII of real people, has the benefit of creating fake data to avoid privacy issues been negated? If so, what level of "real person data matches" would be acceptable, and who would decide it?

Assuming a synthesized body of data contained the PII of "fake" people, and some amount of the records coincidentally matched that of real people, the philosophical question arises as to whether or not the data is really "fake." The question might be best answered (or determined) by the perception of those individuals whose PII coincidentally appeared in the data. If a woman named

Jane A. Thomas born on August 3, 1965 found this data exactly matching her PII in a data base used for data mining for terrorism, would she object to its existence even if told it was completely manufactured by chance?

As mentioned in prior chapters, an alternative to this approach would be to use names that generally do not occur in reality. The abstract range of this data could include semi-realistic names such as "TJamesTTTT Smith" to names comprised of complete gibberish. A problem with this approach, however, is that the greater the extent to which it is desirable that created names do not mirror genuine names/persons, the less a program can rely on existing-name databases for input. Alternatively, at the "gibberish end" of the range of created names, assuming a completely computer-generated process is used to create random strings representing names, the possibility still exists to create names matching those of real persons. Ultimately, any data-synthesis model will most likely include facets/approaches that do not completely mitigate privacy concerns, however minimal.

3. Human/Designer Bias

It is likely that some amount of human bias (or ignorance) will manifest in the system. Humans are always involved in such a process, as a human designed/wrote the program. Again, using the model from Chapter VI, groups of names by ethnicity must be identified and grouped for input. The designer must make some decisions regarding these names.

One example involves nationality versus ethnicity. A surname common in a given geographic area might be labeled by the nationality coinciding with the geographic area, but in fact, the name is considered (by those who possess it) to be belonging to another origin. It is also realistic to assume the designer is not an expert in such matters.

Such errors in explicit or implicit meta-data of the created names (i.e., the manner in which they are treated in the system based on decisions of ethnicity as pointed to by meta-data) could decrease the effectiveness of the data in testing mining programs. For example, the prerequisites of a synthetic body of data might include names/ethnicities to be used as "target names" by the data mining program being tested. The data mining program might also be affected by its own degree of designer bias. In general, a program's ability to function correctly with synthetic data may indicate that the program works, or it may simply indicate an error in the way that the synthetic data was created. A problem may arise if the bias evident in the data/system(s) does not match reality or that of the other.

B. FUTURE WORK

The automated direct-creation model presented in Chapter VI is very basic. This model could be expanded to include more fields representing genuine data bases used in federal data mining programs. Additionally, it is also based on the notion of "seed" records. Automation of data synthesis for full records can only begin once the input records are identified/created, presumably through a mostly manual and time consuming approach. Work to automate the

creation of the seed files would streamline the approach, although it might also offer additional challenges in the areas of meeting goals of data realism.

The automated direct-creation approach was compared with the generative approach in Chapter IV. Of these two non-manual methods, the automated direct-creation is the more simple and easier to implement in terms of design and methodology. The creation of a generative model would be a logical next-step to this area of research, perhaps providing additional insights into the challenges associated with synthetic data. It also offers the potential for more realistic data that better takes into account time as a notion of reality.

A third area of additional research involves the broadening of the understanding of federal data mining programs. This thesis segregated one distinct program based on availability of information and focused on a fraction of its potential data sources. Research providing a more complete framework of existing data mining program architecture, structure, inputs, and methodology would help to illustrate real and additional requirements of useful synthetic data. Once a complete and thorough understanding of a program (or programs) was gained, an attempt to create usable data for testing a specific program could be undertaken.

A fourth area exists in terms of the type(s) of data being synthesized. This thesis focused solely on record-based data. Other interesting data that might be useful to a data mining program could include cell phone tower

records, voice data, or photographs. The input to various data mining programs may consist of other types of data, such as those outlined in Chapter V.

Lastly, extensive knowledge of several real data mining programs could provide a deeper understanding of data used in such programs, to include an analysis of similarities and differences in data. This level of understanding might provide details and characteristics leading to the creation of a more universal body of synthetic data.

C. dONCLUSIONS

This thesis sought to explore and analyze the problem of creating synthetic data for use in data mining for counterterrorism, and to attempt to understand the feasibility of creating a useful body of such data.

One barrier to understanding the extent of this issue remains in the lack of data regarding federal data mining programs. Greater understanding of existing programs would be necessary to better assess the feasibility of creating a useful body of synthetic data.

Using a small sample of records as likely input to federal programs, a simple model to create synthetic records of a small population of persons was created. Creating synthetic data on a vast scale, and which provides both realism and utility, however, would offer a significantly greater challenge, the feasibility of which remains questionable.

APPENDIX A: DHS I-901 FORM

The Department of Homeland Security I-901 (see next page) Form provides a vehicle for would-be non-immigrant students to pay the fee associated with the Student Exchange Visitor Program (SEVP). The form provides ICE with detailed personal information regarding foreign students. This information is a key part in the government's ability to track aliens within the United States, as is included here as an example of source data the government might use as a foundation for data mining programs.

OMB No. 1653-0034
(Expires 09/30/2011)

Department of Homeland Security
U. S. Immigration and Customs Enforcement

I-901, Fee Remittance for Certain F, J and M Nonimmigrants

TYPE OR PRINT IN BLUE OR BLACK INK

1. Last Name *(Surname)*:

2. First Name *(Given Name)*:

3. Middle Name:

WHERE DO YOU WANT YOUR PAYMENT RECEIPT TO BE SENT?

4. Street Address /P.O. Box: Apartment Number:

No. 2 Street Address /P.O. Box:

5. City *(Province)*: **6.** State *(U.S. Address Only)*:

7. Country:

8. Zip Code/Postal Code: **9.** Date of Birth *(mm/dd/yyyy)*: **10.** Gender *(Check one)*: Male: ☐ Female: ☐

11. City *(Province)* of Birth:

12. Country of Birth:

13. Country of Citizenship:

14. School Code *(I-20)* ***(F/M nonimmigrant only):*** 214F OR Program Number *(DS-2019)* ***(J-1 nonimmigrant only):***

15. SEVIS Identification Number: N **16.** Passport Number:

17. Amount to be paid:

A. **F/M only:** ($200) ☐

B. **J-1 only**: Indicate your Exchange Visitor Category *(Check only one of the following boxes)*

Student ($180)	☐	Research Scholar ($180)	☐
Trainee ($180)	☐	Short-term scholar ($180)	☐
Teacher ($180)	☐	Specialist ($180)	☐
Professor ($180)	☐	Intern ($180)	☐
Alien Physician ($180)	☐	Camp Counselor ($35)	☐
Government Visitor ($180)	☐	Summer Work/Travel ($35)	☐
		AuPair ($35)	☐

18. Return Receipt: A. Air Mail ($0) ☐

(You must choose one) B. Expedited Delivery ($35) ☐ Telephone: _____

19. Total amount (add total from 17 and 18): $_____.

Form I-901 (Rev. 10/08)

APPENDIX B: SAMPLE STUDENT STATISTICS (CSU LONG BEACH)

The following list (available at California State University web services) is provided as an example of statistics available for which to provide realism in generating synthetic data. This type of statistics provide a guide for developing non-random parameters within a data generating model, the output of which would ideally mirror reality. The specific data represented here is included as part of the data synthesis exercise involving foreign students and SEVP.

COUNTRY OF CITIZENSHIP FOR INTERNATIONAL STUDENTS

UNDERGRADUATE STUDENTS

Country of Citizenship	2005	2006	2007	2008	2009	2005	2006	2007	2008	2009
Australia	10	6	14	11	18	1.2%	0.7%	1.7%	1.4%	2.5%
Brazil	9	9	8	6	3	1.0%	1.1%	1.0%	0.8%	0.4%
Bulgaria	3	2	1	1	2	0.4%	0.2%	0.1%	0.1%	0.3%
Canada	6	3	8	7	9	0.7%	0.4%	1.0%	0.9%	1.3%
Costa Rica	0	0	1	1	0	0.0%	0.0%	0.1%	0.1%	0.0%
Cyprus	5	5	4	1	0	0.6%	0.6%	0.5%	0.1%	0.0%
France	8	7	6	5	6	0.9%	0.8%	0.7%	0.7%	0.8%
Germany	21	21	11	13	13	2.4%	2.5%	1.3%	1.7%	1.8%
Hong Kong	27	22	21	21	17	3.1%	2.6%	2.5%	2.7%	2.4%
India	10	13	11	6	5	1.2%	1.5%	1.3%	0.8%	0.7%
Indonesia (West Irian)	14	19	20	13	8	1.6%	2.2%	2.4%	1.7%	1.1%
Iran	9	12	8	6	3	1.0%	1.4%	1.0%	0.8%	0.4%
Israel	1	0	0	2	3	0.1%	0.0%	0.0%	0.3%	0.4%
Japan	264	246	241	226	183	30.4%	28.9%	29.1%	29.2%	25.5%
Jordan	4	3	3	2	2	0.5%	0.4%	0.4%	0.3%	0.3%
Kenya	3	2	2	0	1	0.4%	0.2%	0.2%	0.0%	0.1%
Korea, South	109	105	107	101	96	12.6%	12.3%	12.9%	13.0%	13.4%
Lebanon	5	3	0	0	0	0.6%	0.4%	0.0%	0.0%	0.0%
Malaysia	11	4	5	9	6	1.3%	0.5%	0.6%	1.2%	0.8%
Mexico	3	6	5	8	7	0.4%	0.7%	0.6%	1.0%	1.0%
Mongolia	1	2	5	7	6	0.1%	0.2%	0.6%	0.9%	0.8%
Morocco	4	4	3	0	0	0.5%	0.5%	0.4%	0.0%	0.0%
Myanmar (Burma)	6	4	3	1	0	0.7%	0.5%	0.4%	0.1%	0.0%
Nepal	3	5	1	1	2	0.4%	0.6%	0.1%	0.1%	0.3%
Netherlands	2	2	3	3	6	0.2%	0.2%	0.4%	0.4%	0.8%
New Zealand	4	5	2	3	3	0.5%	0.6%	0.2%	0.4%	0.4%
Nigeria	1	0	0	1	1	0.1%	0.0%	0.0%	0.1%	0.1%
Pakistan	13	14	8	5	5	1.5%	1.6%	1.0%	0.7%	0.7%
Peoples Rep of China	15	20	20	20	45	1.7%	2.4%	2.4%	2.6%	6.3%
Peru	3	3	4	3	6	0.4%	0.4%	0.5%	0.4%	0.8%
Philippines	5	5	5	7	9	0.6%	0.6%	0.6%	0.9%	1.3%
Qatar	3	1	5	4	1	0.4%	0.1%	0.6%	0.5%	0.1%
Russian Federation	2	4	4	3	2	0.2%	0.5%	0.5%	0.4%	0.3%
Saudi Arabia	52	74	86	79	73	6.0%	8.7%	10.4%	10.2%	10.2%
Serbia	2	4	6	6	2	0.2%	0.5%	0.7%	0.8%	0.3%
Singapore	4	4	7	4	4	0.5%	0.5%	0.9%	0.5%	0.6%
Sri Lanka (Ceylon)	7	8	9	13	12	0.8%	0.9%	1.1%	1.7%	1.7%
Sweden	13	8	7	8	10	1.5%	0.9%	0.9%	1.0%	1.4%
Syria	5	3	1	0	0	0.6%	0.4%	0.1%	0.0%	0.0%
Taiwan	84	83	62	58	33	9.7%	9.7%	7.5%	7.5%	4.6%
Thailand	9	7	7	5	2	1.0%	0.8%	0.9%	0.7%	0.3%
Turkey	7	4	2	4	3	0.8%	0.5%	0.2%	0.5%	0.4%
United Arab Emirates	23	12	8	10	17	2.7%	1.4%	1.0%	1.3%	2.4%
United Kingdom	10	16	17	14	11	1.2%	1.9%	2.1%	1.8%	1.5%
Vietnam	24	30	31	35	36	2.8%	3.5%	3.7%	4.5%	5.0%
Zambia	1	4	3	2	0	0.1%	0.5%	0.4%	0.3%	0.0%
Other Country	43	38	43	40	46	5.0%	4.5%	5.2%	5.2%	6.4%

See notes on page 4.

APPENDIX C: SAMPLE CALIFORNIA DRIVER'S LICENSE APPLICATION (FORM DL44)

The California driver's license application form is included as a secondary data source for use in the data synthesis exercise. It represents likely data that might be one part of a greater group of data used in a typical data mining program. As part of the data synthesis exercise involving foreign students, it is included as a data source given students' likely desire to operate vehicles while attending university.

HQ MICROGRAPHICS USE ONLY

44

A Public Service Agency

DRIVER LICENSE OR IDENTIFICATION CARD APPLICATION

DO NOT DUPLICATE

1 PURPOSE FOR YOUR VISIT: ✓ the appropriate box(es). ***PRINT USING BLACK OR BLUE INK ONLY***

DRIVER LICENSE (DL)	IDENTIFICATION CARD (ID)	NAME CHANGE/ CORRECTION
☑ Original DL/Permit ☐ Remove Restriction ☐ Renewal ☐ Change/Add Class ☐ Duplicate ____ Lost ____ Stolen ***Complete Parts 2 through 8.***	☐ Original ID Card/Renewal ☐ Senior ID Card/Renewal (Age 62+) ☐ Replacement ____ Lost ____ Stolen ***Complete Parts 2, 3, 5A, 6 & 7 only.***	☐ DL ☐ ID CARD ***Complete Parts 2, 3, 5, 6 & 7 only.***

FOR DMV USE ONLY
BD/LP Code
State/Country
DOCUMENT#
Review: Primary
Secondary Tech ID/Date

2 PLEASE PROVIDE THE FOLLOWING:
NOTE: You must use your true full name. Original documentation may be required. Refer to the *California Driver Handbook.*

Driver License or ID Card Number	State or Country	Expires MO / DAY / YR	Birth Date MO / DAY / YR	Social Security Number
		/ /	10 / 15 / 1989	123 - 45 - 6789

First Name	Middle Name	Last Name	Suffix (Jr., Sr., III)
Joe	Carr	Driver	

Mailing Address, P.O. Box, or Private Mail Box (**Include Box Number, St., Ave., Rd., Blvd., etc.**), Number, Street, Apt/Space No., City, State, Zip Code
123 Main Street Anytown, Ca. 99999

Address Where You Live (**if different from mailing address**), Number, Street, Apt/Space No., City, State, Zip Code

Sex	Hair Color	Eye Color	Height	Weight
☑ M ☐ F	Brown	Brown	6' 1"	165

3 COMPLETE THIS SECTION ONLY IF YOU *ARE NOT* ELIGIBLE FOR A SOCIAL SECURITY NUMBER:

I certify under penalty of perjury under the laws of the State of California that no Social Security Number has ever been issued to me and I am not presently eligible for a Social Security Number. I understand that pursuant to Vehicle Code Section 12801 I must provide my Social Security Number to the Department of Motor Vehicles when one is assigned to me.

Signature	Date
X	

4 LICENSING NEEDS: ✓ the appropriate box(es). Refer to the *California Driver Handbook for additional information.*

BASIC LICENSE	NON-COMMERCIAL LICENSE	AMBULANCE CERTIFICATE
☑ Basic Class C ☐ Motorcycle ***If basic license only, go to Part 5.***	☐ Class A ☐ Class B	☐

5 THE FOLLOWING QUESTIONS MUST BE ANSWERED:

A. Have you applied for a Driver License or Identification Card in California or another state/country using a different name or number within the past ten (10) years? ☐ Yes ☑ No
If yes, print name, DL/ID number, and state or country

B. Have you had your driving privilege or a driver license cancelled, refused, delayed, suspended, or revoked? ☐ Yes ☑ No
If yes, indicate date and reason below.

DATE _____ REASON _____

C. Within the last five years, have you had or experienced any of the medical conditions specified on the back of this form that affects your ability to operate a motor vehicle safely? ***Please read the "Medical Information" on the back of this form before answering.*** ☐ Yes ☑ No
If yes, briefly explain:

6 DO YOU WISH TO REGISTER TO VOTE OR CHANGE POLITICAL AFFILIATION OR VOTER ADDRESS?

DO YOU WISH TO REGISTER TO VOTE OR CHANGE POLITICAL AFFILIATION?	VOTER CHANGE OF ADDRESS
Y ☑ YES—Complete the attached voter form. N ☐ NO—Do not complete attached voter form.	I am a registered voter; I moved and wish to update my voter record. C ☐ to a new county—Complete the attached voter form. S ☐ within the same county—Do not complete the attached form. Your voter record will be automatically updated.

7 DO YOU WISH TO REGISTER TO BE AN ORGAN AND TISSUE DONOR?

DO YOU WISH TO REGISTER TO BE AN ORGAN AND TISSUE DONOR?	
☑ YES! I want to be an organ and tissue donor. ☐ $2 voluntary contribution to support and promote organ and tissue donation.	If you mark "YES!" you will be added to the Donate Life California organ and tissue donor registry and a pink donor dot will be printed on the front of your driver license or identification card. If you are currently registered, you must check "YES!" to have the pink donor dot printed on your license or identification card. If you wish to remove your name from the donor registry, you must contact Donate Life California (see back). The Department of Motor Vehicles can only remove the pink donor dot from your license or identification card.

8 FOR DRIVER UNDER 18, PARENT/GUARDIAN SIGNATURES REQUIRED:
If both parents/guardians have joint custody, ***BOTH MUST SIGN***. I/We accept civil liability for this minor.

Mother's/Guardian's Signature		Date	Daytime Phone Number
X Suzanne Driver		10/17/05	(916) 555-4382
Address *Street*	*Apt No.*	*City*	*State* *Zip*
123 Main St.		Anytown	Ca 99999
Father's/Guardian's Signature		**Date**	**Daytime Phone Number**
X Neal Driver		10/17/05	(916) 555-7205
Address *Street*	*Apt. No.*	*City*	*State* *Zip*
123 Main St.		Anytown	Ca 99999

9 CERTIFICATION: I have read, understand and agree with the contents of this form, including the certifications on the back of this form. ***I certify under penalty of perjury under the laws of the State of California that all the information on this form is true and correct.***

STOP **Do not sign until instructed to do so by a DMV employee.**

Applicant's Signature	FOR DMV FIELD OFFICE USE ONLY
X	
Date / Daytime Phone Number	
(916) 555-4382	

DL 44 (REV 7/2006)

LIST OF REFERENCES

Albrecht, I., Haber, J., & Seidel, H. (2003). Construction of animation of anatomically based human hand models. *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (pp. 98-109). San Diego: Eurographics Association.

Anderson, S. (2003). Total information awareness and beyond: The dangers of using data mining technology to prevent terrorism. Bill of Rights Defense Committee.

Apte, C. (1997). Data Mining: An industrial research perspective. *IEEE Computational Science and Engineering* , 6-9.

Barbaro, M., & T. Zeller, J. (2006, August 9). www.nytimes.com. *A face is exposed for AOL searcher No. 4417749* (Online). New York, New York, United States: New York Times. Retrieved April 11, 2010, from New York Times: http://www.nytimes.com/2006/08/09/technology/09aol.html

Bennett, J., & Lanning, S. (2007). The Netflix prize. *KDD Cup and Workshop.*

Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J., Wang, G., et al. (2003). Crime Data Mining: An overview and case studies. *Proceedings of the 2003 annual national conference on Digital government research* (pp. 1-5). Boston: Digiital Government Society of North America.

CNN Money. (2006, September 26). *CNNMoney.com.* Retrieved April 10, 2010, from CNN.com: http://money.cnn.com/2006/09/26/technology/aol_suit/?postversion=2006092617

Department of Homeland Security. (2009). *Data mining report to Congress.* Department of Homeland Security.

Department of Homeland Security. (2008). *Data mining: Technology and policy.* Department of Homeland Security.

Department of Homeland Security Office of the Inspector General. (2005). *Review of the Immigration and Customs Enforcement's Compliance Enforcement Unit (OIG-05-50).* Washington: Department of Homeland Security.

Department of Homeland Security. (2005). *Privacy impact assessment for the advance passenger information System.* Department of Homeland Security.

Department of Homeland Security. (2007). *Privacy impact assessment update for the advance passenger information system for Customs and Border Protection's general aviation notice of proposed rulemaking.* Department of Homeland Security.

Department of Homeland Security. (2008). *Privacy impact assessment update for the Automated Commercial System (ACS) /* Automated *Commercial Environment (ACE) - Importer Security Filing Data.* Department of Homeland Security.

DeRosa, M. (2004). *Data mining and data analysis for* counterterrorism*.* Washington: Center for Strategic and International Studies.

Dycus, S., Berney, A., Banks, W., & Raven-Hansen, P. (2007). *National security law.* Frederick: Aspen Publishers.

Fayyad, U., & Uthurusamy, R. (2002). Evolving **d**ata **m**ining into **s**olutions for **i**nsights. *Communications of the ACM*, 28-31.

Fienberg, S. (2004). *Homeland insecurity: Datamining, terrorism detection, and confidentiality.* NISS.

Garfinkel, S. (2006). Data surveillance. *IEEE Security and Privacy* , 15-17.

Governmental Accounting Office. (2005). *DATA MINING:* Agencies *have taken key steps to protect privacy in selected efforts, but significant compliance issues remain.* Governmental Accounting Office.

Governmental Accounting Office. (2004). *DATA MINING: Federal efforts cover a wide range of uses.* GAO.

Hoag, J., & Thompson, C. (2007). A parallel general-purpose synthetic data generator. *SIGMOD* , 19-24.

Jeske, D. R., Lin, P., Rendon, C., Xiao, R., & Samadi, B. (2006). Synthetic data generation capabilities for testing data mining tools. *MilCom '06.*

Jeske, D., Samadi, B., Lin, P., Ye, L., Cox, S., Xiao, R., et al. (2005). Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems. *KDD '05* (pp. 756-762). Chicago: ACM.

Jonas, J., & Harper, J. (2006). Effective counterterrorism and the limited role of predictive data mining. *Policy Analysis* .

Lee, W., Stolfo, S., & Mok, K. (1999). A data mining framework for building intrusion detection models. *Proceedings of the 1999 IEEE Symposium on Security and Privacy*, (pp. 120-132).

McHugh, J. (2000). Testing intrusion detection systems; a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln laboratory. *ACM Transactions on Information and System Security*, (pp. 262-294).

Moch, C., & Freiling, F. (2009). The forensic image generator generator (Forensig2). *2009 Fifth Conference on IT Security Incident Management and IT Forensics*, (pp. 78-93). Stuttgart.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large datasets. *IEEE Symposium on Security and Privacy*, (pp. 111-125).

National Commission on Terrorist Attacks Upon the United States. (2004). *The 9/11 Commission report.* Washington: Defense Technical Information Center.

National Research Council. (2008). *Protecting individual privacy in the struggle against terrorists.* The National Academies Press.

Office of the Director of National Intelligence. (2009). *Data mining report.* Office of the Director of National Intelligence.

Popp, R., & Poindexter, J. (2006). Countering terrorism through information and privacy protection technologies. *IEEE Security and Privacy* , 18-27.

Safire, W. (2002, November 14). You are a suspect. *New York Times* .

Waddell, P. (2002). Urban Sim: Modeling urban development for land use, transportation and environmental planning. *Journal of the American Planning Association , 68* (3), 297-314.

Wang, G., Chen, H., & Atabakhsh, H. (2004). Automatically detecting deceptive criminal identities. *Communications of the ACM* , 71-76.

Weizenbaum, J. (1983). ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the ACM* , 23-28.

Zdanowicz, J. (2004). Detecting money laundering and terrorist financing via data mining. *Communications of the ACM* , 53-55.

Zhao, W., Chellappa, R., Phillips, P., & Rosenfeld, A. (2003, December). Face recognition: a literature survey. *ACM Computing Surveys* , pp. 339-458.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Marine Corps Representative
   Naval Postgraduate School
   Monterey, California

4. Director, Training and Education, MCCDC, Code C46
   Quantico, Virginia

5. Director, Marine Corps Research Center, MCCDC, Code C40RC
   Quantico, Virginia

6. Marine Corps Tactical Systems Support Activity (Attn: Operations Officer)
   Camp Pendleton, California